


머신 러닝 알고리즘을 이용한 COVID-19 Risk 분석 및 Safe Activity 지원 시스템

COVID-19 Risk Analytics and Safe Activity Assistant System with Machine Learning Algorithms


요약

최근 COVID-19 으로 인한 팬데믹 상황이 이어지면서 전세계적으로 이전의 일반적인 생활이 불가능할 뿐만 아니라 많은 사람들이 사망하고 있다. 더불어 COVID-19 에 대한 백신 개발마저 어렵고 효능 또한 보장하기 어려운 상황이다. 따라서, 사람들은 뉴노멀 (New Normal) 시대를 준비해가며 COVID-19 에서 안전할 수 있는 생활방식을 추구하는 추세이다. 하지만 아직까지 COVID-19 의 위험도를 수치적으로 나타내거나 안전한 활동 등에 대한 정보를 제공하는 연구가 부족하다. 따라서 본 논문에서는 머신러닝 알고리즘 기반 COVID-19 의 위험도 분석과 안전 활동에 대한 정보 제공에 대한 방법을 제안한다. 본 논문에서는 COVID-19 에 대해 영향을 받을 수 있는 요소들을 이용한 Risk 분석 메트릭을 제안한다. 해당 메트릭을 통해 개인 및 그룹에 대한 위험도를 수치적으로 나타낼 수 있다. 또한 지역 및 사람에 대한 특성 등을 통해 클러스터링 알고리즘에 적용하여 COVID-19 에 대한 위험도를 제공한다. 해당 클러스터링 결과를 활용하여 사용자가 안전한 생활을 할 수 있는 데 도움이 되는 정보를 제공하는 것을 보였다.

 주제어: COVID-19, 머신 러닝, 클러스터링, Risk 분석, Safe Activity 지원

Abstract

Lately, as the pandemic situation caused by COVID-19 continues, it is hard to return a normal life, and many people have been dying globally. In addition, most research of COVID-19's vaccine announce that not only it is hard to develop the vaccine but also the efficacy of the vaccine might be not perfect to eradicate COVID-19. As a result, people prepare new normal life as changing their lifestyle for safety. However, there are few the researches showing the disease's risk as numerically and providing the information for safety activity. Therefore, we propose machine learning algorithm based COVID-19 analytics and safe activity assistance system. In this paper, we propose a risk analysis metrics using factors that can be affected for COVID-19. Risks for individuals and groups can be represented numerically through the corresponding metrics. Also, it applies to clustering algorithms through characteristics of regions and people, and provides a risk for COVID-19. By using the results of the clustering, it was shown that information is provided to help users live a safe life.

 keyword : COVID-19, Machine Learning, Clustering, Risk Analytics, Safe Activity Assistance

1. 서론

COVID-19 가 중국 우한에서 처음 보고된 이후 국제 사회는 인류 역사에서 겪어보지 못한 새로운 시대를 경험하고 있다. COVID-19 는 치사율은 높진 않지만 강한 전염성이 특징으로 짧은 시간에 전 세계적으로 전파되어 팬데믹(Pandemic) 사태를 불러일으켰다. 또한 COVID-19 에 전염되면 사망하기까지 평균 한달도 걸리지 않지만 COVID-19 을 위한 백신은 출시하기까지 수개월도 더 남은

실정이며 그 효과에 대한 보증은 더욱 요원한 상황이다.

이미 COVID-19 가 온 사회에 만연하고 치료책이 없는 이 시점에서 인류 및 지역 사회에 요구되는 것은 지역 사회 구성원 스스로, COVID-19 의 전염을 최대한 늦추기 위해 감염원에서 멀어지는 방어적인 행동에 임하는 것이다. 방어적인 행동에는 경제활동 중단부터 시작하여 외출금지, 야외활동 금지 등의 자가격리 또는 사람 간의 접촉 및 스킨십 금지 등이 있다.

하지만 현실적으로 COVID-19 의 백신이 나올

때까지 일련의 방어적인 행동들을 시행하는 건 불가능에 가깝다. 또한 이러한 방어적인 행동들을 팬데믹 초기엔 잘 지켜지는 모습을 보였지만 시간이 지날수록 방어적인 행동들에 피로감을 보이고 COVID-19 의 전염성 및 위험성에 무디어지게 되어 기본적인 안전 수칙조차 지키지 않는 현상이 발생하고 있다.

이러한 현상을 해결하기 위한 관련 연구로는 COVID-19 의 리스크에 관한 연구 [1-4], COVID-19 감염 경로 예측 및 분석에 관한 연구 [5-15], COVID-19 팬데믹 시대와 지역사회에 대한 연구 [16-22], 인적데이터를 클러스터링화하여 분석하는 연구 [23-26] 등이 있다. 하지만 이 연구들은 위험도를 수치적으로 표현하거나 안전한 활동에 대한 정보를 제공하는 연구는 매우 적게 진행되고 있다.

본 논문에선 이에 대한 해결책으로 COVID-19 데이터를 이용한 메트릭 및 클러스터링을 통해 COVID-19 에 대한 위험성을 계산하고 수치화하여 개인에게 막연했던 COVID-19 의 위험수치를 보여주어 방어적 행동을 유발하거나 지속시키는 방식을 제안한다. 이 해결책은 COVID-19 데이터로 산출된 개인의 주변 환경 및 사람들에 대한 정보를 보여줌으로써 개인 혹은 주변 환경 이 어느 수준의 위험인지 보여준다.

본 논문에서는 COVID-19 리스크 평가 체계를 요소 및 메트릭으로 제안하고, 인구 및 그룹 기반 클러스터링 기법을 제안한다. 그리고 Safe Activity Assistant를 통해 현 COVID-19 팬데믹 사태 및 향후 전염병 위기를 위한 해결책을 제시한다.

2. 관련 연구

본 장에서는 COVID-19 Risk와 이에 대한 분석 그리고 팬데믹 하의 Safe Activity에 관련된 최근 연구들을 살펴본다.

우선 COVID-19 Risk를 탐지하거나 감염 경로

및 분석에 관한 연구에 대해 살펴본다. Alaa의 연구는 인구 밀도를 분석하는 방법으로 COVID-19 가 확산될 위험이 있는 구역을 식별하는 전략을 제시하였다 [27]. 하지만 고위험 구역을 분석하는 데 그치고 본 전략을 지역 사회에 도움이 되기 위한 해결책을 제시하지 못했다.

Wang의 연구는 지역 사회, 특히 도시 시스템에서 COVID-19 의 출현과 확산을 모델링하고 분석하는 연구를 진행하였으며 COVID-19 의 확산의 원인요소를 규명하고자 하였다 [28]. 하지만 위험요소를 활용하여 COVID-19 의 확산을 통제하는 방안에 대해서는 제시하지 않았다.

Lee의 연구는 COVID-19 가 한국 내 지역사회에서 미친 영향을 조사하고 당면한 현안과제를 분석한 뒤 향후 발전 방향을 제안하였다 [29]. 하지만 관공서 및 언론사에 대한 제언과 민간에도 이에 연관된 협력에 대한 요구에 그치고 민간의 역할 및 가능성에 대해서는 제시하지 않았다.

마지막으로 Song의 연구[30]는 대화형 웹 기반 매핑 플랫폼을 통해 사회적 거리두기를 강화할 수 있는 방안을 제시하였다. 하지만 지역민들의 경각심을 고취시키는 방안이나 지역 내부 또는 지역 사회 간 움직임을 자제하는 방안까지는 제시하지 않고 있다.

현재 COVID-19 전염성과 관련되어 COVID-19 를 치료 및 예방하기 위한 연구, 감염을 억제하거나 통제하기 위한 제안 및 연구가 활발하게 이루어지고 있다. 하지만 대다수의 연구들은 COVID-19 전염성에 대해 통제, 예방 혹은 억제 등 방식에 대해 일면만을 다루거나 이를 위한 IT 기술 혹은 관공서의 역할군에 대해서만 강조하고 있다. COVID-19 에 대해 피로감을 느끼고 둔감해지는 지역 사회 시민들의 경각심을 다시금 고취시키기 위한 방안은 제시되지 않았다. 즉, COVID-19 와 직접 마주하는 대중의 지속적인 안전을 위한 명시적이고 수치화된 연구는 아직 활발하지

않다.

3. COVID-19 리스크 메트릭

본 절에서는 COVID-19 리스크 평가에 대한 설명과 필요성 그리고 COVID-19 리스크 평가를 위한 요소들과 메트릭을 소개한다.

3.1 COVID-19 리스크 정의

COVID-19 리스크 평가는 COVID-19 데이터를 이용하여 감염자(COVID-19에 걸린 확진자), 접촉자(감염자와 접촉한 사람) 혹은 장소에 대한 위험성을 계산하고 수치화하여 등급을 산정하는 것을 의미한다. 이 평가의 최종목적은 COVID-19 전염성 및 위험성에 대해 둔감해지는 사회 구성원들에게 경각심을 고취시키는 것에 있다. 즉, 전염성 및 위험성을 수치화하여 더 방어적인 행동을 유발하고 사회안전망에 기여하도록 돕는데 있다.

COVID-19 리스크 평가는 두 가지 방식으로 산출할 수 있다. GSR(Group Safety Risk)과 ISR(Individual Safety Risk)이다. 그리고 이 두 방식에 공통적으로 들어가는 요소인 Severity(위험도)가 필요하다. 이번 장의 3.2 절, 3.3 절 항목에서 GSR과 ISR을 산출하기 위한 요소 및 메트릭을 제안한다.

COVID-19 리스크 평가를 계산하기 위해 필요한 요소는 크게 감염자나 접촉자의 Severity 계산을 위한 요소, GSR 값 계산을 위한 요소 그리고 ISR값 계산을 위한 요소로 나뉜다.

3.2 COVID-19 그룹 리스크 평가 메트릭

GSR에서 그룹은 시, 군, 구처럼 행정구역을 의미하는 것뿐만 아니라 건물 혹은 편의시설처럼 국소적인 장소를 의미한다. 즉 GSR은 해당 그룹이 얼마나 위험한 상태인지 보여주는 척도다. GSR값의 범위는 [0, 1]이며 0이면 무결한 장소를 의미하고 1이면 감염성이 매우 높은 위험한 상태를 의미한다.

이러한 GSR값을 산출하기 위한 요소들의 집합을 식 (1)과 같이 나타낼 수 있다.

$$GSRFactorSet_{GroupName} = \{GSR_{element1}, GSR_{element2}, \dots, GSR_{elementN}\} \quad (1)$$

COVID-19의 위험성과 감염성에 대한 연구는 현재 진행형이므로 GSR값 산출을 위한 요소는 추가 및 제거될 수 있다. 예를 들어 한 그룹의 GSR값을 구하는데 필요한 요소가 해당 그룹의 면적 및 그룹에 머물렀던 사람들의 정보일 경우, $GSRFactorSet_{GroupName}$ 을 식 (2)과 같이 나타낼 수 있다.

$$GSRFactorSet_{GroupName} = \{SeveritySet, Safe\} \quad (2)$$

$SeveritySet$ 은 그룹, 즉 지역, 건물 혹은 편의시설에 있었던 사람들의 위험도를 수치화한 값들의 집합을 나타낸다. 즉 식 (3)과 같이 나타낼 수 있다.

$$SeveritySet = \{severity_{y1}, severity_{y2}, \dots, severity_{yN}\} \quad (3)$$

A severity value of a infected or contacted person}

$SeveritySet$ 의 원소인 $severity_i$ 의 범위는 (0, 1]이며 0에 가까울수록 건강한 상태를 의미하며 1에 가까울수록 그 사람이 위험하다는 것을 의미한다. Severity는 감염자 혹은 접촉자가 얼마나 위험한지 보여주는 척도이다. 이러한 Severity값을 산출하기 위한 요소들은 식 (4)와 같다.

$$SeverityFactorSet_{personID} = \{sElement1, sElement2, \dots, sElementN\} \quad (4)$$

COVID-19의 위험성과 감염성에 대한 연구는 현재 진행형이므로 $SeverityFactorSet_{personID}$ 의 요소는 추가 및 제거될 수 있다. $sElement$ 은 감염되거나 감염자와 접촉하고 경과한 날, 나이 등이 있다. 접촉 후 경과 일자만 적용하였을 때 Severity값을 산출하기 위한 요소를 식 (5)과 같이 나타낼 수 있다.

$$SeverityFactorSet_{personID} = \{DaysAfterInc\} \quad (5)$$

$DaysAfterIncurred$ 는 사람이 감염되거나 감염자와 접촉한 후 경과한 날을 의미하며 정수로 표시된다.

$Severity_{personID}$ 값을 산출하는 메트릭에는 감염자에 대한 메트릭과 접촉자에 대한

메트릭이 있다. 먼저 감염자에 대한 메트릭은 식 (6)과 같다.

$$Severity_{personID} = 1 - (DaysAfterInfection) \times 0.05 \quad (6)$$

이때 $DaysAfterInfection$ 은 양의 정수이며 범위는 $[0, 14]$ 이다. 15 째일부터는 $Severity_{personID}$ 는 0으로 간주된다.

접촉자에 대한 메트릭은 식 (7)과 같다.

$$Severity_{personID} = 1 - DaysAfterContact \times 0.05 \times 0.5 \quad (7)$$

이때 $DaysAfterContact$ 는 양의 정수이며 범위는 $[0, 14]$ 이다. 15 일째부터는 감염자의 경우와 마찬가지로 0으로 간주된다. 예를 들어 감염자가 감염된 후 5일이 경과한 후에 $Severity_{personID}$ 를 측정한다면 0.75가 나온다.

현재 메트릭은 감염 혹은 접촉 후 경과한 날짜만 고려한 메트릭으로, $SeverityFactorSet_{personID}$ 에 요소가 추가됨에 따라 메트릭에 변동사항이 있거나 추가적인 연산이 있을 수 있다. 예를 들어 Age가 $SeverityFactorSet_{personID}$ 의 원소로 추가된다면 $Severity_{personID}$ 를 구하는 메트릭은 식 (8)과 같다.

$$Severity_{personID} = SeverityAge_{personID} \quad (8)$$

이때 $SeverityAge_{personID}$ 값과 $SeverityDays_{personID}$ 값의 범위는 $(0, 1]$ 이므로 $Severity_{personID}$ 값의 범위도 $(0, 1]$ 이 된다. 예를 들어 나이만 다르고 동일한 조건의 사람 둘이 동시에 $Severity$ 값을 측정하게 된다면 40살의 $SeverityAge_{personID}$ 는 20살의 $SeverityAge_{personID}$ 값보다 높게 나오므로 40살의 최종 $Severity_{personID}$ 값이 더 크게 나올 것이다.

$SafetyDensity$ 는 해당 지역, 건물 혹은 편의시설의 내부 사람들의 밀집도를 나타낸다. 범위는 $[0, 1]$ 이며 0에 가까울수록 공간이 넓거나 사람이 적음을 의미하고 1에 가까울수록 비좁은 공간에 사람이 많다는 것을 의미한다.

예를 들어 20명 정도가 적정인 피트니스 센터에 40명이 있고, 그 중 감염자가 2명이라면

$SeveritySet$ 은 $\{0.8, 0.7\}$ 로 측정될 수 있고 $SafetyDensity$ 는 1에 근접하게 나온다.

GSR값을 산출하기 위해서 $SeveritySet$ 과 $SafetyDensity$ 의 계산이 선행되어야 한다. 먼저 $SeveritySet$ 의 평균을 구한다. $SeveritySet$ 평균은 식 (9)과 같이 나타낼 수 있다.

$$AvgSeverity = avg(SeveritySet) \quad (9)$$

그리고 $SafetyDensity$ 를 산출하는 공식은 식 (10)과 같이 나타낼 수 있다.

$$SafetyDensity = \frac{1}{min(1, PeopleN \times MinSafeArea_{Group})} \quad (10)$$

$PeopleN$ 은 해당 그룹에 있었던 사람 수를 의미한다. $MinSafeArea$ 는 COVID-19의 전염을 막기 위해 사람과 사람 사이에 유지되어야 하는 거리를 기반으로 계산한 면적을 의미한다. 이 거리는 COVID-19 연구에 따라 유동적으로 변할 수 있다. $GroupTotalArea$ 는 해당 그룹의 총 면적을 의미한다. $min()$ 함수는 입력 값 중에 최소값을 출력하는 함수이다. 즉, 식 (10)에서 $min()$ 함수와 정수 1로 $SafetyDensity$ 의 최대값은 1이 된다.

식 (9)와 (10)을 기반으로 $AvgSeverity$ 와 $SafetyDensity$ 를 기반으로 GSR값을 구하는 공식은 식 (11)과 같다.

$$GSR_{GroupName} = AvgSeverity \times SafetyDensity \quad (11)$$

$AvgSeverity$ 의 범위는 $[0, 1]$ 이고 $SafetyDensity$ 의 범위가 $[0, 1]$ 이므로 $GSR_{GroupName}$ 의 범위는 $[0, 1]$ 이다. 0에 가까울수록 COVID-19의 위험성이 낮고 무결한 장소를 의미하며 1에 가까울수록 고위험군이 있거나 많은 감염자들이 존재했음을 의미한다.

예를 들어 20명 정도가 적정인 피트니스 센터에 40명이 있고, 그 중 감염자가 2명이라면 $SeveritySet$ 은 $\{0.8, 0.7\}$ 이고 식 (9)에 의해 $AvgSeverity$ 는 0.75가 나온다. 그리고 센터의 $GroupTotalArea$ 가 200이고 $MinSafeArea$ 가 10이라면 식 (10)에 의하여 연산 결과는 20이 나오게 되고 $SafetyDensity$ 는 $min()$ 함수에 의해 1이 된다.

3.3 COVID-19 개인 리스크 평가 메트릭

다음으로 ISR은 개인이 현재 얼마나 위험한 상태인지 보여주는 척도다. 결과값의 범위는 [0, 1]이며 0 이면 무결한 상태를 의미하고 1 이면 매우 위험한 상태를 의미한다. ISR은 감염된 상태의 위험도를 의미하는 Severity와 달리 주변 환경을 고려하여 현재 개인에게 얼마나 큰 위험이 있는지를 의미한다. ISR을 산출하는 데 필요한 요소들은 식 (12)와 같다.

$$ISR_{FactorSet}^{personID} = \frac{ISR_{Element1} + ISR_{Element2}}{2} \quad (12)$$

COVID-19의 위험성과 감염성에 대한 연구는 현재 진행형이므로 ISR값 산출을 위한 요소는 추가 및 제거될 수 있다. ISR을 산출하기 위한 요소로는 개인의 Severity, 주변 그룹에 대한 정보, 그리고 주변 그룹과 연관성을 나타내는 정보 등이 될 수 있다. 이러한 요소들을 가지는 $ISR_{FactorSet}^{personID}$ 를 식 (13)과 같이 나타낼 수 있다.

$$ISR_{FactorSet}^{personID} = \frac{Severity_{personID} + \frac{1}{n} \sum_{i=1}^n SurroundingGSRSet_i^{personID}}{2} \quad (13)$$

$Severity_{personID}$ 는 사용자(ISR값을 측정하는 사람)의 Severity를 의미한다. $Severity_{personID}$ 의 범위는 [0, 1]이며 0 이면 건강한 상태를 의미하며 1 에 가까울수록 그 사람이 위험하다는 것을 의미한다. $RelatedGroupSet_{personID}$ 는 개인과 연관성이 깊은 그룹의 집합을 의미하며 원소인 $rPlace$ 는 구체적인 특정 그룹을 의미한다. $RelatedGroupSet_{personID}$ 는 식 (14)와 같이 나타낼 수 있다.

$$RelatedGroupSet = \{rPlace \mid A \text{ related group}\} \quad (14)$$

예를 들어 술집의 코로나 확산은 현지 직장인보다 현지 중고등학생에게 더 적게 영향을 끼칠 것이며, 대구의 코로나 확산은 태백의 시민보다 서울 시민에게 더 민감하게 작용할 것이다.

$SurroundingGSRSet_{personID}$ 은 개인의 주변 환경에 위치한 GSR의 집합을 의미하며 원소인 $GSR_{GroupName}$ 은 $GroupName$ 의 GSR값을 의미한다. $Surrounding-GSRSet_{personID}$ 은 식 (15)와 같이

나타낼 수 있다.

$$SurroundingGSRSet_{personID} = \{GSR_{GroupName} \mid GSR_{GroupName} \text{ is a related group}\} \quad (15)$$

가능한 $GroupName$ 의 범위는 거주지 주변 환경과 현재 위치한 장소의 주변 환경 정보를 포함한다. 예를 들어 $GroupName$ 은 거주지 근처 상점의 상호나 학교부지 전체를 나타낼 수 있다.

ISR값을 산출하기 위해 $RelatedGroupSet_{personID}$ 과 $SurroundingGSRSet_{personID}$ 을 통해 유의미한 GSR을 분류하는 작업이 선행되어야 한다. 개인에게 적합한 그룹을 추출하기 위한 집합을 구하기 위한 공식은 식 (16)과 같다.

$$TargetGSRSet_{personID} = \{tGSR \mid tGSR \text{ is an element of } SurroundingGSRSet_{personID}\} \quad (16)$$

$TargetGSRSet_{personID}$ 는 주변 그룹 중에서도 사용자와 연관성이 있는 그룹을 $tGSR$ 로 추출하여 원소로 가지고 있는 집합을 의미한다. $tGSR$ 은 GSR과 동일하며 범위와 의미 또한 동일하다.

이러한 $TargetGSRSet_{personID}$ 을 바탕으로 산출되는 ISR값은 식 (17)과 같다.

$$ISR_{personID} = \frac{w_s \cdot Severity_{personID} + \frac{1}{n} \sum_{i=1}^n LivingDistance_i - tGSRDistance_i}{w_s + w_g} \quad (17)$$

w_s 와 w_g 는 각각 Severity에 대한 가중치와 $tGSR$ 에 대한 가중치를 의미하며 두 가중치의 합은 1 이고 Severity와 $tGSR$ 이 ISR값에 미치는 영향을 조절한다. $Severity_{personID}$ 는 ISR을 사용자의 Severity값을 의미한다. $LivingDistance$ 는 사용자의 생활 반경을 의미한다. $tGSRDistance$ 는 사용자와 $tGSR$ 과의 거리를 의미한다. $LivingDistance$ 와 $tGSRDistance$ 로 사용자로부터 거리에 따른 가중치를 조절할 수 있다. n 은 $tGSR$ 의 수, 즉 $TargetGSRSet$ 의 원소의 수를 의미한다. $tGSR$ 은 $TargetGSRSet$ 의 원소로 GSR과 범위와 의미가 동일하다. 이 요소들을 계산해서 나온 ISR값의 범위는 [0, 1]이며 0 에 가까울수록 사용자가 COVID-19 의 감염 위험에서 안전하다는 것을 의미하며 1 에 가까울수록 생활 반경 내에 COVID-19 감염 위험이 많다는 것을 의미한다.

예를 들어 *Severity*에 더 큰 영향력을 주고자 하면 w_s 에 0.7 을, w_g 에 0.3 을 부여한다. 사용자가 *SeverityIncurredDate*만 이용할 경우, 감염된 후 6 일이 지났다고 가정하면 사용자의 *SeverityPersonID*는 0.7 이 된다. 주변에 *tGSR*이 많고 그 거리가 대부분 멀지 않을 경우 시그마항의 값은 1 에 가깝게 나올 것이다. 그러면 주어진 식에 따라 $ISR_{PersonID}$ 는 1 에 가까운 값이 도출된다.

4. 클러스터링 기반 COVID-19 에 대한 안전활동 알고리즘 및 활용

본 장에서는 3 절에서 언급한 요소들과 메트릭을 통해 나온 결과값으로 클러스터링을 진행한다. 그리고 클러스터링에 사용하는 거리 함수를 일반적으로 사용되는 거리 함수와 다르게 가중치를 적용한 논리적 거리 함수를 설명한다.

4.1 클러스터링 알고리즘

클러스터링이란 거리 함수와 데이터들의 특징을 이용하여 데이터들 간의 거리를 계산하고 이를 기반으로 여러 개의 클러스터를 생성하여 데이터를 클러스터별로 분류하는 것을 의미한다.

가중치를 적용한 논리적 거리 함수에서 가중치는 클러스터링에 사용되는 특징들에 적용되며 각 특징에 영향력을 조절하게 된다. 또한 가중치의 합은 1 로 유지되어 가중치 간의 비율을 유지하여 일부 특징이 과도하게 영향력을 가지게 되는 것을 방지한다.

예를 들어 유클리드 거리 함수를 이용하는 경우, 특징 *Severity*, *Age* 대한 점 $p(p_{severity}, p_{age})$ 과 $q(q_{severity}, q_{age})$ 이 있고 두 점사이의 거리를 $distance_{pq}$ 라 하면 $distance_{pq}$ 의 값을 구하는 공식은 식 (18)과 같다.

$$distance_{pq} = \sqrt{p_{severity}^2 - q_{severity}^2} \quad (18)$$

이때 $p_{severity}$ 와 p_{age} 는 어떤 데이터 p 의

*Severity*값과 *Age*값을 의미하며 $q_{severity}$ 은 q_{age} 는 p 와 다른 데이터 q 의 *Severity*값과 *Age*값을 의미한다.

그리고 가중치 *Severity*에 대한 가중치 $w_{severity}$ 와 *Age*에 대한 가중치 w_{age} 를 추가한 변형 공식은 식 (19)와 같다.

$$distance_{pq} = \sqrt{w_{severity}^2 p_{severity}^2 - q_{severity}^2} \quad (19)$$

이런 거리함수를 이용한 클러스터링의 기본알고리즘은 (표 1) 기본적인 클러스터링 알고리즘

(Table 1 과 같다.

(표 1) 기본적인 클러스터링 알고리즘

(Table 1) Algorithm of Fundamental Clustering

Algorithm 1. Algorithm of Creating Cluster

Input: algoName: name of algorithm, cN: number of clusters, df: distance function method, wList: weight list of features

Output: clusterInstacnce : instance of cluster algorithm

1	function createCluster(algoName, cN, df, wList):
2	newCluster := selectCluster(algoName)
3	newCluster.setFunction(df)
4	if cN inputted:
5	newCluster.setClusterNum(cN)
6	end if
7	If wList inputted:
8	newCluster.setWeight(wList)
9	endif
10	return clusterInstacnce

함수에 클러스터 알고리즘 이름, 클러스터의 수, 거리 함수, 가중치 리스트를 전달한다 (Line 1). selectCluster() 함수에 클러스터 알고리즘 이름을 전달하여 사용할 클러스터를 결정한다 (Line 2). 클러스터 알고리즘에 클러스터의 수를 입력해야 할 경우 클러스터의 수를 setClusterNum을 통해 클러스터 인스턴스에 전달한다 (Line 4~6). 그리고 거리 함수에 적용할 가중치 리스트를 입력했다면 setWeight() 함수를 통해 클러스터 인스턴스에 전달한다 (Line 7~9). 그 후, 클러스터 인스턴스를 반환한다 (Line 10).

4.2 그룹 기반 클러스터링 알고리즘

그룹 기반 클러스터링이란 사람을 이용하여

클러스터링하는 것을 의미한다. 본 절에서는 그룹 기반 클러스터링의 필요성, 알고리즘 및 활용 방안에 대해 설명한다.

4.2.1 그룹 기반 클러스터링의 필요성

현재 CDC(Centers for Disease Control and Prevention, 미국의 질병관리예방본부)에 따르면 COVID-19의 전염 방식은 사람과 사람사이에서 호흡기 비말을 통해 전파되며 공기 중 전파는 밝혀진 감염경로가 아니다 [31]. 하지만 WHO(World Health Organization)에 따르면 공기 중 전파의 가능성을 완전히 배제할 수는 없으며 특정 조건 하에서는 공기로 통한 전염이 충분히 가능하다고 경고하며 엘리베이터 같은 좁은 실내에서 짧은 시간내에 감염되는 사례도 보고되고 있다 [32].

그렇기에 COVID-19의 영향력이 사람으로부터 나오기도 하지만 감염자나 접촉자가 머문 장소로부터 감염될 수 있기 때문에 감염자나 접촉자에 대한 분류만 아니라 이들이 머무르거나 지나간 장소에 대한 분류에 대한 요구도 필연적이라 할 수 있다. 예를 들어 술집에서 감염자와 술을 마신 접촉자가 지하철을 타고 편의점을 방문한 후 집에 도착했다면, 해당 술집과 지하철 그리고 편의점에서 추가적인 감염자가 발생할 수 있다. 그러므로 이런 그룹들은 즉시 방역대상이 포함되며 방역 전까지는 다른 무결한 그룹들과 다르게 관리되어야 할 것이다. 그룹 기반 클러스터링은 이러한 경우에 핵심적인 역할을 할 것이다.

4.2.2 그룹 기반 클러스터링 알고리즘

인구 기반 클러스터링과 마찬가지로 그룹 기반 클러스터링 알고리즘도 특징이 요구되며 사용자는 원하는 종류를 선택할 수 있다. 예를 들어 필수적으로 포함되는 GSR값을 비롯하여 BusinessCategory, DailyDensity, RelatedGroup 등을 선택하여 클러스터링을 진행할 수 있다. 예를 들어 특징으로 ISR값과 BusinessCategory를

사용할 경우 클러스터링 알고리즘의 입력 데이터는 다음과 같다.

(표 2) 그룹 클러스터링 입력 데이터 예시

(Table 2) Example of Group Clustering Input Data

ID	GSR	BusinessCategory
1	0.33	Bar
2	0.89	Bar
3	0.78	Restaurant
4	0.11	Laundry
5	0.01	Apartment

입력 데이터에는 다섯개의 데이터가 있으며 각 데이터는 식별을 위해 ID 값을 가지고 있다. 그리고 각 데이터는 GSR값과 BusinessCategory값을 가지고 있다. GSR값의 범위는 [0, 1]이며 0에 가까울수록 COVID-19로부터 무결한 그룹을 의미하며 1에 가까울수록 COVID-19의 전염성 위험이 높은 그룹임을 의미한다. BusinessCategory은 업종을 의미한다.

그리고 4.1의 가중치를 적용한 거리 함수를 이용하여 클러스터링을 진행한다. 이때 거리 함수 또한 유클리드 거리를 비롯하여 원하는 거리 함수를 선택할 수 있으며 가중치 또한 입력 가능하다. 이러한 과정을 통해 나오는 출력데이터 예시는 다음과 같다.

(표 3) 그룹 클러스터링 출력 데이터 예시

(Table 3) Example of Group Clustering Output Data

ID	GSR	BusinessCategory	ClusterID
1	0.33	Bar	1
2	0.89	Bar	2
3	0.09	Restaurant	3
4	0.11	Laundry	3
5	0.01	Apartment	3

출력 데이터에는 ClusterID라는 새로운 열이 추가되었다. 해당 열은 그룹 기반 클러스터링을 통해 분류된 클러스터의 ID이다. ID 3, 4, 5는 비교적 낮은 GSR값으로 3으로 분류되었으며 ID 1, 2은 GSR값이 높은 편이지만 카테고리가 다르기 때문에 ID 1은 ClusterID 1로 분류되었고 ID 2은 ClusterID 2로 분류되었다.

4.3 인구 기반 클러스터링 알고리즘

인구 기반 클러스터링이란 사람을 이용하여 클러스터링하는 것을 의미한다. 본 절에서는 인구 기반 클러스터링의 필요성, 알고리즘 및 활용 방안에 대해 설명한다.

4.3.1 인구 기반 클러스터링의 필요성

지금은 COVID-19 에 의한 팬데믹 시대로 지금껏 인류가 경험하지 못한 새로운 시대이다. 팬데믹 시대에는 1 명의 감염자 혹은 접촉자는 슈퍼 감염자가 될 수 있는 위험한 가능성이 있으며 이에 따른 전염병을 통제하는 방법이 요구되는 만큼, 감염자 및 접촉자를 분류하는 방법도 필연적으로 요구된다. 이때 감염자와 접촉자와 가까운 사람들의 생활 습관 및 행동은 무결한 곳에 사는 사람들에 비해 더 방어적으로 임할 필요가 있다. 인구 기반 클러스터링을 통한 사람 분류는 지역 사회 주민들의 방어적인 행동에 일조하고 지역 전염율을 낮추는데 도움을 줄 수 있다. 예를 들어 이태원 클럽에서 감염자가 존재해서 수많은 접촉자가 생겼을 경우, 이태원 지역 주민 및 접촉자들의 거주지 지역 주민들의 행동은 더욱 조심스러워야 한다. ISR을 통한 인구 기반 클러스터링은 지역 주민들의 방어적인 행동에 일조하여 COVID-19 의 지역사회 전파율을 감소시킬 것이다.

4.3.2 인구 기반 클러스터링 알고리즘

클러스터링을 진행하기 위해서는 데이터의 특징이 요구된다. 마찬가지로 인구 기반 클러스터링 알고리즘도 특징이 필요하다는 건 동일하지만 사용자는 원하는 종류를 선택할 수 있다. 예를 들어 필수적으로 포함되는 ISR값을 비롯하여 Age, Gender, Related Disease, Address 등을 선택하여 클러스터링을 진행할 수 있다.

예를 들어 특징으로 ISR값과 Age를 사용할 경우 클러스터링 알고리즘의 입력 데이터는 다음과 같다.

(표 4) 사람 클러스터링 입력 데이터 예시

(Table 4) Example of People Clustering Input Data

ID	ISR	Age
----	-----	-----

1	0.01	27
2	0.05	43
3	0.03	71
4	0.65	22
5	0.89	22

입력 데이터에는 다섯개의 데이터가 있으며 각 데이터는 식별을 위해 ID 값을 가지고 있다. 그리고 각 데이터는 ISR값과 Age값을 가지고 있다. ISR값의 범위는 [0, 1]이며 0 에 가까울수록 COVID-19 로부터 안전한 환경을 의미하며 1 에 가까울수록 COVID-19 로부터 전염될 위험이 높은 환경에 있음을 의미한다. Age는 0 이상의 정수로 표시되며 해당 데이터의 나이를 의미한다.

해당 데이터에 4.1 절의 가중치를 적용한 거리 함수를 이용하여 클러스터링을 진행한다. 이때 거리 함수 또한 유클리드 거리를 비롯하여 원하는 거리 함수를 선택할 수 있으며 가중치 또한 입력 가능하다. 이러한 과정을 통해 나오는 출력 데이터 예시는 다음과 같다.

(표 5) 사람 클러스터링 출력 데이터 예시

(Table 5) Example of People Clustering Output Data

ID	ISR	Age	ClusterID
1	0.01	27	1
2	0.05	43	1
3	0.03	71	2
4	0.65	22	3
5	0.89	22	3

출력 데이터에는 ClusterID라는 새로운 열이 추가되었다. 해당 열은 인구 기반 클러스터링을 통해 분류된 클러스터의 ID이다. ID 4, 5 는 비교적 높은 ISR 값으로 3으로 분류되었으며 ID 1, 2, 3 은 ISR값이 낮지만 Age의 편차가 커서 ID 1, 2 는 ClusterID 1 로 분류되었고 ID 3 은 ClusterID 2 로 분류되었다.

4.4 리스크 기반 클러스터링의 활용

본 절에서는 인구 기반 및 그룹 기반 클러스터링 알고리즘의 활용 방안에 대해서 제안한다. 각 클러스터링의 경우 ISR값 또는 GSR값과 특징 타입에 따라 사람 혹은 그룹들이 클러스터별로 분류된다. 이 클러스터들은 ISR값

또는 GSR값과 선택한 특징에 따라 현재 각 클러스터들에 속한 사람이나 그룹들이 COVID-19에 대해 얼마나 위험한지 가능할 수 있으며, 미래에 각 클러스터들이 COVID-19에 대해 얼마나 취약해질 수 있는 지를 예측하는데 활용될 수 있다. 다만 그룹 기반 클러스터링은 인구 기반 클러스터링과 다르게 방역시스템에 의해 즉시 COVID-19로부터 무결함을 의미하는 값인 0으로 초기화 될 수 있다.

4.4.1 소규모 시설 단위로 위험도별 그룹 분류

그룹 기반 클러스터링은 COVID-19에 전염될 위험이 높은 그룹과 그렇지 않은 그룹을 나누는데 목적이 있으며 기본적인 활용방안은 이러한 목적에 기인한다. 본 방안을 위한 classifyFacilityGroup 메소드의 구성은 (표 6) 소규모 시설 그룹 분류를 위한 알고리즘 (Table 6와 같다).

(표 6) 소규모 시설 그룹 분류를 위한 알고리즘

(Table 6) Algorithm of classifying facility group

Algorithm 2. Algorithm to classify facility group

Input: ipd: infected people data, cpd: contacted people data, fgd: facility group data, cluster: instance of cluster

Output: clusterResult: cluster id list including input data, clusterInfo: concrete information of cluster

1	function classifyFacilityGroup(ipd, cpd, fgd, cluster):
2	integratedData := incorporateData(ipd, cpd, fgd)
3	clusterIDList := cluster(integratedData)
4	clusterResult := incorporateData(integratedData, clusterIDList)
5	clusterInfo := extractClusterInfo(clusterResult)
6	return clusterResult, clusterInfo

함수에 입력된 감염자 데이터, 접촉자 데이터, 그리고 소규모 시설 그룹 데이터를 incorporateData() 함수로 통합한다 (Line 2). 통합된 데이터로 클러스터링 후 결과로 나온 Cluster ID 리스트까지 통합한다 (Line 3~4). 통합 데이터를 이용하여 클러스터링을 통한 클러스터들의 정보를 추출한다 (Line 5). 그 후, 통합 데이터와 클러스터 정보를 반환한다 (Line

6).

(표 6) 소규모 시설 그룹 분류를 위한 알고리즘 (Table 6의 알고리즘을 기반으로 한 클러스터링의 예를 들면, 편의시설이나 교육시설 같은 건물 단위로 그룹 단위로 설정하고 대구의 한 교회에서 대규모 집단 감염이 발생했다고 가정한다. 그렇다면 해당 교회 및 주변 지역 일대 그룹들의 GSR이 매우 높은 값을 가지게 될 것이며 해당 그룹들의 클러스터군은 고위험의 클러스터로 분류될 것이다. 이 수치들은 대구 주민들의 ISR값 갱신 및 주민들의 클러스터에 큰 영향을 미치며 대구 주민들의 행동에 큰 영향을 준다. 감염자 및 접촉자들이 분리되고 지역 일대에 방역이 마무리되면 GSR값은 크게 감소되고 지역 주민들의 방어적인 행동은 조금 완화될 수 있을 것이다. 다만 지역 내 자가 격리된 접촉자들이 여전히 존재하니 ISR은 0이 되지 않으며 지역 주민들에게 COVID-19 확산에 대해 지속적인 경계심을 줄 수 있다. 이로 인해 지역 사회의 방어적인 행동의 지속성을 증가를 기대할 수 있다.

4.4.2 행정 구역 단위로 위험도별 그룹 분류

그룹 기반 클러스터링에서 그룹은 건물 단위가 아닌 행정 구역 단위로 설정하여 어느 행정구역이 COVID-19 감염성이 높은 지역인지 확인할 수 있다. 이를 위한 classifyProvinceGroup 메소드 및 알고리즘은 (표 7) 행정구역 그룹 분류를 위한 알고리즘 (Table 7와 같다).

(표 7) 행정구역 그룹 분류를 위한 알고리즘

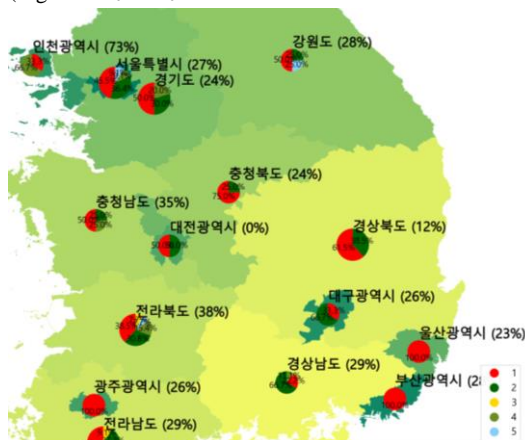
(Table 7) Algorithm of classifying province group

Algorithm 3. Algorithm to classify high risk province group**Input:** pd: people data, pgd: province group data, cluster: instance of cluster**Output:** clusterResult: cluster id list including input data, clusterInfo: concrete information of cluster

1	function classifyFacilityGroup(pd, pgd, cluster):
2	integratedData := incorporateData(pd, pgd)
3	clusterResult := cluster(integratedData)
4	clusterResult := incorporateData(integratedData, clusterResult)
5	clusterInfo := extractClusterInfo(clusterResult)
6	return clusterResult, clusterInfo

함수에 입력된 감염자, 접촉자 그리고 무결한 자들의데이터와 행정 구역 그룹 데이터를 incorporateData() 함수로 통합한다 (Line 2). 통합된 데이터로 클러스터링 후 결과로 나온 Cluster ID 리스트까지 통합한다 (Line 3~4). 통합 데이터를 이용하여 클러스터링을 통한 클러스터들의 정보를 추출한다 (Line 5). 그 후, 통합 데이터와 클러스터 정보를 반환한다 (Line 6).

(표 7) 행정구역 그룹 분류를 위한 알고리즘 (Table 7) 의 알고리즘을 기반으로 예를 들면 임의로 100 명의 데이터를 생성하여 클러스터링을 진행하였으며 그 결과는 아래의 (그림 1) 큰 행정구역 그룹 별 클러스터링 결과 (Figure 1 과 같다.



(그림 1) 큰 행정구역 그룹 별 클러스터링 결과

(Figure 1) Result of Clustering by Province Group

클러스터링은 makeLargeGroupCluster

메소드에서 k-means 알고리즘으로 5 개의 클러스터링으로 초기 세팅하였으며 각 특징에 동일한 가중치를 부여한 유클리드 거리함수를 이용하였다. 100 명의 데이터는 감염 날짜, Covid-19 Status(감염, 접촉, 무결), 거주지 데이터, 면적 데이터 및 이동경로 데이터를 가지고 있다. 또한 ISR 및 GSR 계산을 통해 각 데이터는 ISR값 및 GSR값 또한 포함하고 있다.

(그림 1) 큰 행정구역 그룹 별 클러스터링 결과 (Figure 1) 의 파이 그래프의 크기는 해당 그룹의 속한 인적 데이터의 수를 의미한다. 원의 면적은 5 개의 클러스터에 포함되는 사람에 따라 나뉜다. 행정구역의 색은 행정구역마다 ID가 있어 행정구역의 구분을 위해 조금씩 다르다. 그룹 명에 포함된 퍼센티지는 전체 인적 데이터 중에 감염자 혹은 접촉자 수를 의미한다.

이러한 데이터로 고위험 그룹의 방문을 자제하여 COVID-19 확산을 억제하는 행동을 유발하는 효과를 기대 할 수 있다. 예를 들어 경상남도에 거주하는 지역 주민에게 부산이나 대구의 방문을 스스로 자제하게 만드는 효과를 기대할 수 있다.

4.4.3 업종 단위로 위험도별 그룹 분류

그룹 기반 클러스터링을 이용하여 소규모 단위 그룹들을 업종 별로 분류하여 연관성이 높은 그룹들끼리 클러스터를 형성할 수 있다. 이를 위한 classify-RelatedGroup 메소드 및 알고리즘은 (표 8) 관련 그룹을 분류하기 위한 알고리즘

(Table 8 과 같다.

(표 8) 관련 그룹을 분류하기 위한 알고리즘

(Table 8) Algorithm to classify related group

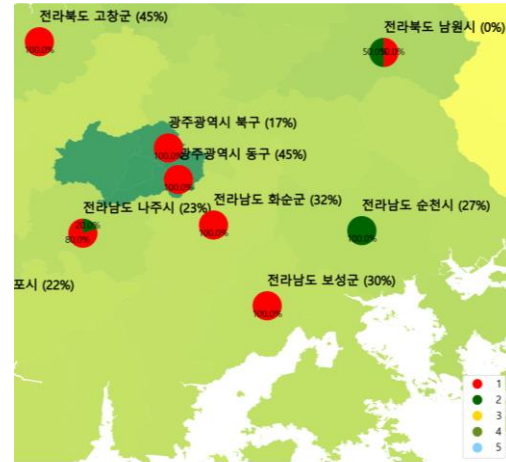
Algorithm 4. Algorithm to classify related group**Input:** ipd: infected people data, cpd: contacted people data, sgd: specific group data, cluster: instance of cluster**Output:** clusterResult: cluster id list including input data, clusterInfo: concrete information of cluster

1	function classifyFacilityGroup(ipd, cpd, sgd, cluster):
2	integratedData := incorporateData(ipd,

	cpd, sgd)
3	clusterResult := cluster(integratedData)
4	clusterResult := incorporateData(integratedData, clusterResult)
5	clusterInfo := extractClusterInfo(clusterResult)
6	return clusterResult, clusterInfo

함수에 입력된 감염자 데이터, 접촉자 데이터, 그리고 업종 단위 그룹 데이터를 incorporateData() 함수로 통합한다 (Line 2). 통합된 데이터로 클러스터링 후 결과로 나온 Cluster ID 리스트까지 통합한다 (Line 3~4). 통합 데이터를 이용하여 클러스터링을 통한 클러스터들의 정보를 추출한다 (Line 5). 그 후, 통합 데이터와 클러스터 정보를 반환한다 (Line 6).

(표 8) 관련 그룹을 분류하기 위한 알고리즘 (Table 8을 기반으로 한 클러스터링의 예를 들면, 학생과 연관도가 높은 그룹은 식당, 오락시설, 교육시설, 독서실 등이 될 수 있다. 그리고 술집, 관공서, 유흥시설 등의 그룹들과 연관성은 아주 낮을 것이다. 이렇게 업종별로 클러스터링을 진행하면 해당 업종이 다른 업종에 비해 얼마나 위험한지 알 수 있다. 예를 위하여 전라남도 지방의 문화 센터에 견학하고자 하는 선생님이 그룹 기반 클러스터링을 이용하여 안전한 문화센터를 찾는다고 가정하고 클러스터링을 진행하였으며 그 결과는 (그림 2) 작은 행정구역 그룹 별 클러스터링 결과 (Figure 2와 같다. (그림 1) 큰 행정구역 그룹 별 클러스터링 결과 (Figure 1과 다르게 상세히 표시하기 위해 전라남도 지역 중심으로 표시하였다.



(그림 2) 작은 행정구역 그룹 별 클러스터링 결과
(Figure 2) Result of Clustering by Small Province Group

클러스터링은 makeSmallGroupCluster 메소드에서 k-means 알고리즘으로 5 개의 클러스터링으로 초기 세팅하였으며 각 특징에 동일한 가중치를 부여한 유클리드 거리함수를 이용하였다. 30 곳의 데이터는 감염자 방문 날짜, 소독 여부, 주소 데이터 및 면적 데이터를 가지고 있다. 또한 GSR 계산을 통해 각 데이터는 GSR값 또한 포함하고 있다. 그리고 GSR값 계산을 위해 임의의 100 명의 무작위 데이터를 생성하였다. 마지막으로 지역의 보다 자세한 표현을 위해 (그림 2) 작은 행정구역 그룹 별 클러스터링 결과 (Figure 2에는 전라남도과 전라북도 일부만을 표현하였다.

(그림 2) 작은 행정구역 그룹 별 클러스터링 결과

(Figure 2의 파이 그래프의 크기는 해당 그룹의 속한 인적 데이터의 수를 의미한다. 원의 면적은 5 개의 클러스터에 포함되는 사람에 따라 나뉜다. 행정구역의 색은 행정구역마다 ID가 있어 행정구역의 구분을 위해 조금씩 다르다. 그룹 명에 포함된 퍼센티지는 특정 기준 이상의 Severity값을 가진 사람들의 비율을 의미한다. 여기서는 0.2로 설정하였다.

이러한 그룹 기반 클러스터링에 의하면 클러스터링 ID가 1에 가까울수록 0에 가까운

GSR을 가진 그룹이고 5에 가까울수록 높은 GSR을 가진 그룹이다. 즉, 이 경우에는 순천시, 남원시 그리고 나주시 문화센터는 피하는 것이 좋으며 클러스터 1에 속한 그룹 중에서도 감염자 비율 퍼센티지가 낮은 광주광역시 북구의 문화센터가 제일 안전할 수 있다.

이러한 방식으로 그룹 기반 클러스터링을 이용하여 특정 업종의 그룹이 얼마나 위험한지 파악할 수 있다. 행정 구역별 고위험 그룹 분류와 더불어 사용하면 특정 행정 구역에 사는 사람은 어떤 업종의 방문 및 시설이용을 자제해야 하는지 파악할 수 있으며 이로 인해 COVID-19의 확산을 낮추는 행동을 기대할 수 있다.

4.4.4 선택적 고위험 그룹 분류

인구 기반 클러스터링은 지역 사회의 구성원이 개인의 안전을 위해 활용할 수 있으며 이는 인구 기반 클러스터링의 목적에 해당한다. 또한 지역 사회 구성원은 자신의 목적에 맞게 ISR과 다른 특징에 가중치를 부여하여 클러스터링을 할 수 있다. 이를 위한 classify-IndivisualStatus 메소드 및 알고리즘은 과 같다.

(표 9) ISR 값 데이터로 분류하기 위한 알고리즘

(Table 9) Algorithm to classify by ISR value data

Algorithm 5. Algorithm to classify by ISR value data

Input: wList: weight list of features, algoName: name of algorithm, cN: number of clusters, df: distance function method, ipd: infected people data, cpd: contacted people data, upd: user personal data, gd: group data including GSR, weightedCluster: cluster instance which has weight list

Output: clusterResult: cluster id list including input data, clusterInfo: concrete information of cluster

1	function classifyIndividualStatus(wList,
	algoName, cN, df, ipd, cpd, upd, gd):
2	weightedCluster = createCluster(wList, algoName, cN, df)
3	integratedData := incorporateData(ipd, cpd, upd, gd)
4	clusterResult := weightedCluster(integratedData)
5	clusterResult := incorporateData(integratedData, clusterReslut)

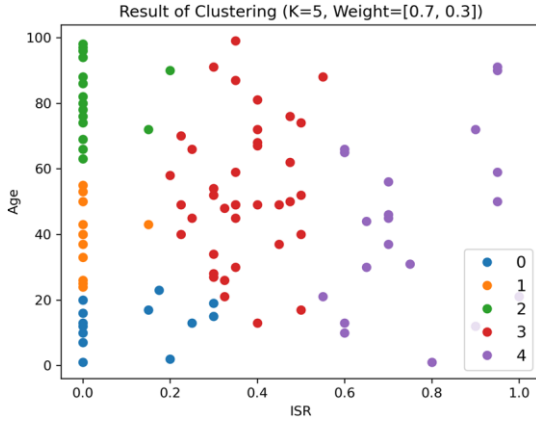
6	clusterInfo := extractClusterInfo(clusterResult)
7	return clusterResult, clusterInfo

가중치 리스트, 클러스터 알고리즘 이름, 클러스터의 수, 그리고 거리 함수를 매개변수로 전달하여 가중치가 부여된 클러스터 인스턴스를 생성한다 (Line 2). 함수에 입력된 감염자 데이터, 접촉자 데이터, 그리고 소규모 시설 그룹 데이터를 incorporateData() 함수로 통합한다 (Line 3). 통합된 데이터로 클러스터링 후 결과로 나온 Cluster ID 리스트까지 통합한다 (Line 4~5). 통합 데이터를 이용하여 클러스터링을 통한 클러스터들의 정보를 추출한다 (Line 6). 그 후, 통합 데이터와 클러스터 정보를 반환한다 (Line 7).

(표 9) ISR 값 데이터로 분류하기 위한 알고리즘 (Table 9를 기반으로 한 클러스터링의 예를 들면, 서울시 관악구에 감염자 3명이 원인을 알 수 없는 경로로 등장했을 경우, 감염자들과 함께 있던 사람들은 즉시 접촉자가 된다. 그리고 감염자들과 접촉자들의 Severity가 높은 값으로 갱신되며 주변 사람들의 ISR과 주변 그룹들의 GSR이 갱신된다.

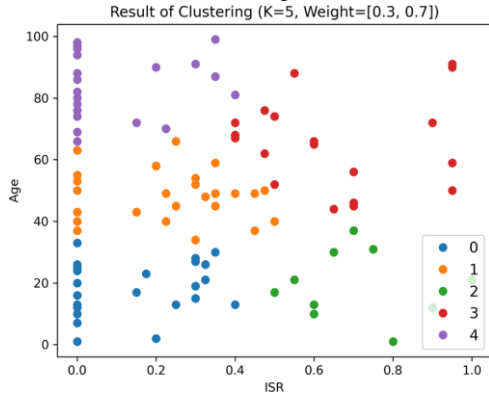
이러한 과정에서 추출된 데이터가 100명의 데이터라고 하고, 그 중 ISR에 가중치를 두고 가중치를 진행할 경우 그 결과는 (그림 3) 가중치를 ISR에 0.7, 나이에 0.3을 준 사람 클러스터링

(Figure 3과 같고 나이에 가중치를 두고 클러스터링을 진행한 결과는 (그림 4) 가중치를 ISR에 0.3, 나이에 0.7을 준 사람 클러스터링 (Figure 4와 같다.



(그림 3) 가중치를 ISR에 0.7, 나이에 0.3 을 준 사람 클러스터링

(Figure 3) People Clustering as weight 0.7 to ISR and 0.3 to Age



(그림 4) 가중치를 ISR에 0.3, 나이에 0.7 을 준 사람 클러스터링

(Figure 4) People Clustering as weight 0.3 to ISR and 0.7 to Age

(그림 3) 가중치를 ISR에 0.7, 나이에 0.3 을 준 사람 클러스터링

(Figure 3)에서는 ISR에 Age보다 큰 가중치를 부여했더니 클러스터링이 ISR에 민감하게 클러스터링을 구성한 모습을 보여주었고, 반대로 (그림 4) 가중치를 ISR에 0.3, 나이에 0.7 을 준 사람 클러스터링 (Figure 4)에서는 Age에 민감하게 클러스터링을 구성한 모습을 보여주었다.

이렇게 개인의 클러스터가 변하고 고위험의 클러스터에 속하게 될 경우 개인의 COVID-19의 감염율을 높이게 되는 고위험 클러스터로 갱신되었으므로 사회 활동의 감소를 기대할 수

있다. 그리고 클러스터의 변동으로 현재 감염자 및 접촉자와 이들로 인한 GSR값의 변동으로 현재 개인의 포지션이 어떻게 변하며 어떤 위험에 새롭게 마주했는지 한번에 인지시켜 줄 수 있다. 즉, 텍스트 문자 및 감염자 경로 공개 등의 조치와 달리 COVID-19의 위험성을 수치화하여 사회 구성원의 막연했던 경각심의 재고를 기대할 수 있다.

4.4.5 그룹 변동에 따른 갱신될 위험 그룹 예측

COVID-19 관련 데이터를 이용한 그룹 기반 클러스터링과 인구 기반 클러스터링의 특징을 이용하여 지역 사회 구성원의 미래 행동에 대한 결과를 예측하여 보여줄 수 있다. 이를 위한 predictISRValue 메소드 및 알고리즘은 (표 10)

다른 그룹으로 이동할 때 ISR을 예측하기 위한 알고리즘

(Table 10)과 같다.

(표 10) 다른 그룹으로 이동할 때 ISR을 예측하기 위한 알고리즘

(Table 10) Algorithm to predict ISR when moving to another group

Algorithm 6. Algorithm to predict ISR when moving to another group

Input: uISR: user ISR data, rgd: current group data including GSR around user residence, vg: group data to visit including GSR

Output: fISR: predicted future ISR value, fUserClusterID: predicted future user cluster ID, fGroupClusterIDList: changed groups of future cluster ID of the group user inhabits

1	function predictFValues():
2	fISR := computeFISR(vg)
3	fUserClusterID := computeFClusterID(fISR, rgd)
4	for aGroup in range number of vg:
5	fGCID := computeFGCID(fISR, aGroup)
6	if fGCID \neq current GroupClusterID
7	fGroupClusterIDList \leftarrow (aGroup, fGCID)
8	endif
9	endfor
10	return fISR, fUserClusterID, fGroupClusterIDList

함수에 개인의 ISR 데이터, 개인 거주지 주변의 그룹데이터, 그리고 방문할 곳의 그룹데이터를 입력한다 (Line 1). 한 개인이 방문할 그룹의 GSR에 따라 갱신될 ISR을

계산한다 (Line 2). 변경된 ISR 값에 따라 갱신될 개인의 클러스터 ID를 계산한다 (Line 3). 변경된 ISR값 및 개인의 클러스터에 따라 갱신되는 개인 거주지 주변의 그룹 정보와 클러스터 ID를 추출한다 (Line 4~9). 이때 변경되지 않는 그룹의 정보는 추출하지 않는다 (Line 6). 그 후, 갱신될 개인의 ISR값, 클러스터 ID, 그리고 거주지 주변 그룹의 클러스터 ID 리스트를 반환한다 (Line 10).

예를 들어 낮은 ISR을 가진 사람이 높은 GSR 그룹들이 포진한 곳에 방문할 경우, 사용자는 본 지역 방문을 통해 변경될 본인의 ISR값을 받아볼 수 있다. 그리고 GSR와 주변 지역 주민들의 ISR값을 이용하여 그룹 기반 클러스터링을 진행할 경우 갱신된 ISR값에 따라 주변 그룹들의 클러스터링 군이 고위험 군으로 변동될 수 있다. 즉, COVID-19의 팬데믹 하에서 자신의 행동이 가져올 수 있는 변화를 수치적으로 명시할 수 있다.

이러한 방식으로 사용자의 행동이 지역 사회에 끼치는 영향을 시각적으로 명확히 보여줌으로써 COVID-19 확산에 대한 방어적인 행동을 유발하거나 지속할 수 있는 효과 등을 기대할 수 있다.

5. 결 론

전세계적인 COVID-19의 확산으로 인해 많은 사람들의 삶의 질이 저하되었을 뿐만 아니라 목숨까지 위험한 상황에 놓여있다. COVID-19라는 질병의 특징 상, 질병에 걸린 후 완치가 된다 해도 후유증 등의 여파가 있을 것이라는 연구 결과들이 나오고 있다. 따라서 사람들이 현재 환경에서 COVID-19의 위험에서 벗어나 안전한 환경을 추구하고 있다. 하지만, 사람들에게 직접적인 수치로 정보를 제공하는 연구와 실험은 매우 적게 이루어지고 있다.

본 논문에서는 COVID-19에 대해 영향을 미치는 요소들을 이용하여 수치적으로

위험도를 나타내는 메트릭들을 제시하였다. 해당 메트릭들을 통해 주변의 건물 뿐 만 아니라 도시, 국가 등의 단위에서 해당 지역의 위험도와 한 개인의 처해진 상황과 해당 사람의 정보를 기반으로 한 위험도를 계산할 수 있음을 보였다.

더불어 클러스터링 알고리즘을 통해 COVID-19의 위험도를 인구 기반과 그룹 기반의 클러스터링 방식을 제안하였다. 또한 이를 이용하여 사람들에게 직접적으로 영향을 줄 수 있는 활용 방안 5가지를 제시하였다.

따라서, 본 논문의 연구를 통해 실제 사람들에게 COVID-19의 위험에 대한 경각심을 주고 위험도를 평가하고 안전한 곳을 찾고자 하는 사람들의 요구를 만족시킬 수 있게 된다.

References

- [1] G. Stewart, K. Heusden and G. A. Dumont, "How control theory can help us control Covid-19," *IEEE Spectrum*, Vol.57, No.6, pp.22-29, June 2020.
- [2] M. Jain, P. K. Bhati, et al., "Modelling Logistic Growth Model for COVID-19 Pandemic in India," in *Proceedings of 5th International Conference on Communication and Electronics Systems (ICCES 2020)*, Coimbatore, India, pp.784-789, July 2020.
- [3] B. Wang, Y. Sun, et al., "Risk-Aware Identification of Highly Suspected COVID-19 Cases in Social IoT: A Joint Graph Theory and Reinforcement Learning Approach," *IEEE Access*, Vol.8, pp.115655-115661, June 2020.
- [4] Pakpour, A.H. and Griffiths, M.D., "The fear of COVID-19 and its role in preventive behaviors," *Journal of Concurrent Disorders*, Vol.2, No.1, pp.58-63, April 2020.
- [5] V. Chamola, V. Hassija, et al., "A Comprehensive Review of the COVID-19 Pandemic and the Role of IoT, Drones, AI, Blockchain, and 5G in Managing its Impact," *IEEE Access*, Vol.8, pp.90225-90265, May 2020.
- [6] M. Abdel-Basset, R. Mohamed, et al., "A Hybrid COVID-19 Detection Model Using an Improved Marine Predators Algorithm and a Ranking-Based Diversity Reduction Strategy," *IEEE Access*, Vol.8, pp.79521-79540, April 2020.
- [7] A. K. Nandi, "Data Modeling With Polynomial Representations and Autoregressive Time-Series Representations, and Their Connections," *IEEE Access*, Vol.8, pp.110412-110424, June 2020.
- [8] E. Hernández-Orallo, P. Manzoni, et al., "Evaluating How Smartphone Contact Tracing

- Technology Can Reduce the Spread of Infectious Diseases: The Case of COVID-19," *IEEE Access*, Vol.8, pp.99083-99097, May 2020.
- [9] E. Montes-Orozco et al., "Identification of COVID-19 Spreaders Using Multiplex Networks Approach," *IEEE Access*, Vol.8, pp.122874-122883, July 2020.
- [10] M. Small and D. Cavanagh, "Modelling Strong Control Measures for Epidemic Propagation With Networks—A COVID-19 Case Study," *IEEE Access*, Vol.8, pp.109719-109731, June 2020.
- [11] S. S. Arun and G. Neelakanta Iyer, "On the Analysis of COVID19 - Novel Corona Viral Disease Pandemic Spread Data Using Machine Learning Techniques," *4th International Conference on Intelligent Computing and Control Systems (ICICCS 2020)*, Madurai, India, pp.1222-1227, May 2020.
- [12] N. Zheng et al., "Predicting COVID-19 in China Using Hybrid AI Model," *IEEE Transactions on Cybernetics*, Vol.50, No.7, pp.2891-2904, July 2020.
- [13] Wei Xu and Chongyang Chen, "Research on the Influencing Factors and Management Countermeasures of College Students' Sense of Security under the Environment of Big Data-an Empirical Analysis based on the Event of COVID-19," in *Proceedings of the 2020 The 3rd International Conference on Big Data and Education (ICBDE 2020)*, Machinery, New York, USA, p21-25, April 2020.
- [14] K. El Emam, "Seven Ways to Evaluate the Utility of Synthetic Data," *IEEE Security & Privacy*, Vol.18, No.4, pp.56-59, July 2020.
- [15] Soon Ae Chun, Alen Chih-Yuan Li, et al., "Tracking Citizen's Concerns during COVID-19 Pandemic" in *Proceedings of 21st Annual International Conference on Digital Government Research (dg.o 2020)*, New York, USA, p322-323, June 2020.
- [16] A. Ashok, M. Guruprasad, et al., "A Machine Learning Approach for Disease Surveillance and Visualization using Twitter Data," in *Proceedings of International Conference on Computational Intelligence in Data Science (ICCIDS 2019)*, Chennai, India, pp.1-6, Feb. 2019.
- [17] R. B. Duffey and E. Zio, "Analysing Recovery From Pandemics by Learning Theory: The Case of CoVid-19," *IEEE Access*, Vol.8, pp.110789-110795, June 2020.
- [18] Andrea Remuzzi, Giuseppe Remuzzi, "COVID-19 and Italy: what next?," *The Lancet Haematology*, Vol.7, No.5, pp.1225-1228, May 2020.
- [19] P. Staszkievicz, I. Chomiak-Orsa and I. Staszkievicz, "Dynamics of the COVID-19 Contagion and Mortality: Country Factors, Social Media, and Market Response Evidence From a Global Panel Analysis," *IEEE Access*, Vol.8, pp.106009-106022, June 2020.
- [20] A. Khattar, P. R. Jain and S. M. K. Quadri, "Effects of the Disastrous Pandemic COVID 19 on Learning Styles, Activities and Mental Health of Young Indian Students - A Machine Learning Approach," in *Proceedings of 4th International Conference on Intelligent Computing and Control Systems (ICICCS 2020)*, Madurai, India, pp.1190-1195, May 2020.
- [21] R. F. Sear et al., "Quantifying COVID-19 Content in the Online Health Opinion War Using Machine Learning," *IEEE Access*, Vol.8, pp.91886-91893, May 2020.
- [22] S. Greenstein, "Uncomfortable Economic Waters," *IEEE Micro*, Vol.40, No.4, pp.134-136, July. 2020.
- [23] El-Atem N, Irvine KM, Valery PC, et al. "Identifying areas of need relative to liver disease: geographic clustering within a health service district," *Australian Health Review*, Vol.41, No.4, pp.407-418, Aug. 2017.
- [24] Poon, A. F. Y., "Impacts and shortcomings of genetic clustering methods for infectious disease outbreaks," *Virus Evolution*, Vol.2, No.2, pp.1-9, July 2016.
- [25] Athanasios Tsanas and Siddharth Arora, "Large-scale Clustering of People Diagnosed with Parkinson's Disease using Acoustic Analysis of Sustained Vowels: Findings in the Parkinson's Voice Initiative Study," in *Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies*, Valletta, Malta, pp.369-376, Jan. 2020.
- [26] L. Li, S. Xu, S. Wang and X. Ma, "The Diseases Clustering for Multi-source Medical Sets," in *Proceedings of International Conference on Identification, Information and Knowledge in the Internet of Things (IIKI 2016)*, Beijing, China, pp.294-298, Oct. 2016.
- [27] A. A. R. Alsaedy and E. K. P. Chong, "Detecting Regions At Risk for Spreading COVID-19 Using Existing Cellular Wireless Network Functionalities," *IEEE Open Journal of Engineering in Medicine and Biology*, Vol.1, pp.187-189, June 2020.
- [28] B. Wang, S. Xu and M. Mansouri, "Modeling the emergence of COVID-19: a systems approach," in *Proceedings of IEEE 15th International Conference of System of Systems Engineering (SoSE)*, Budapest, Hungary, pp.445-450, June 2020.
- [29] Moo-Sik Lee, "Overcoming the COVID-19 Epidemics with Communities in Korea," *Journal of agricultural medicine and community health*, Vol.45, No.1, pp.41-46, March 2020.
- [30] Song Gao, Jinmeng Rao, et al., "Mapping county-level mobility pattern changes in the United States in response to COVID-19," *SIGSPATIAL Special*, Vol.12, No.1, pp.16-26, March 2020.
- [31] "How to Protect Yourself & Others | CDC," Centers for Disease Control and Prevention. last modified April 24, 2020, accessed July 21, 2020, <https://www.cdc.gov/coronavirus/2019->

머신 러닝 알고리즘을 이용한 COVID-19 Risk 분석 및 Safe Activity 지원 시스템

- ncov/prevent-getting-sick/prevention.html
- [32] "Modes of transmission of virus causing COVID-19: implications for IPC precaution recommendations," World Health Organization. last modified 29 March, 2020, accessed July 21, 2020, <https://www.who.int/news-room/commentaries/detail/modes-of-transmission-of-virus-causing-covid-19-implications-for-ipc-precaution-recommendations>